Introduction¹

This book rests on the assumption that an in-depth analysis of the insanity defense (including the fundamental question of whether it is properly regarded as a legal defense) cannot be made without tackling the problems of philosophical anthropology, i.e., without asking some fundamental metaphysical questions about the nature of human agency or, more generally, about the nature of human beings.² Certainly, not all important questions that can be posed about the insanity defense will require forays into philosophical anthropology, but, as will be demonstrated in this book, the questions which do require such a treatment are numerous. Therefore, at the outset I deem it necessary to present the philosophical-anthropological view (picture) of human nature that will determine (more or less directly) my solutions to the questions that arise in the context of the insanity defense. However, prior to presenting this view, it is expedient to make some general remarks about the subject of my analysis: the insanity defense.

Along with automatism, involuntary intoxication, ignorance, entrapment, or duress, the insanity defense belongs to a specific group of legal defenses, viz. to excuses, which imply that even though the actions of an agent are considered to be bad, the agent himself – due to some fact about his mental state – cannot be held responsible for this act.³ The group of excuses differs from the second group of legal defenses – that of justifications (including, e.g., self-defense or necessity), which imply that an agent can be held responsible

¹ The work on this publication was partly financed by the Priority Research Area Society of the Future under the program “Excellence Initiative – Research University” at the Jagiellonian University in Krakow.

² This kind of approach is by no means widespread in books devoted to the insanity defense. One title among the few exceptions is Michael Moore’s excellent *Law and Psychiatry: Rethinking the Relationship* (1984). The structure of his argumentation (the defense of the legal view of persons as rational and autonomous against some psychiatric ideas: that badness is illness, that our mental life is ruled by the unconscious, and that the unity of self is a fiction) is very different from the one adopted in this book (although his “legal view of persons” is close to the personalist view defended in the present book).

³ In the case of entrapment and duress, strictly speaking this is not some fact about the agent’s mental state that excludes his responsibility, but some fact about the situation in which he found himself that made it extremely difficult for him to avoid acting wrongfully; but, in general, in the case of excuses, it is the mental state that counts.

for his *prima facie* bad act, but, due to some special circumstances of the act, this act can be justified, and therefore, on deeper reflection, cannot be regarded as bad (immoral/illegal) – indeed, it has become permitted or even obligatory.⁴ Two additional remarks are in order here. First, I shall use the term “insanity defense” throughout this book even though it is hard to deny that, given the semantic associations of the adjective “insane,” it may have some stigmatizing element. A possible alternative could be “the mental nonresponsibility defense.” Yet in spite of its advantages: it is not stigmatizing and eliminates “the confusion as to whether retarded defendants who are not mentally ill are entitled to assert the defense” (Ellis, Luckasson 1985: 441), this felicitous term is still not well established in the relevant literature, so I shall stay with the traditional “insanity defense.”⁵ Secondly, the insanity defense should be distinguished from the mental competence required to stand trial. The insanity defense refers to the defendant’s mental state at the time the crime was committed. If the defense is successful, then the defendant is found not guilty. By contrast, competence (or incompetence) to stand trial refers to the mental state of the defendant at the time when the trial is supposed to take place. If the defendant is found to be incompetent, then the trial is delayed until he regains his competence. In general, for a defendant to be deemed competent, he must be able to understand the charges against him and to assist in his own defense. The two institutions are obviously intricately connected with each other but my analysis will focus solely on the insanity defense, making only occasional references to the institution of competence to stand trial.

Let me now pass on to a presentation of the philosophical-anthropological view underlying my examination of the insanity defense. I will not provide its justification in a systematic and in-depth manner as this would require the writing of a very different book; I will just state it explicitly and present several arguments in its favor. While I believe it to be correct, needless to say I am fully aware that many scholars are not inclined to share my opinion. Those who reject it may interpret my analyses as a kind of philosophical exercise aimed at delineating the consequences that a given picture of human nature has for the insanity defense. This kind of interpretation should be facilitated by the fact that my analyses will be pursued in a broader context, since I strive

⁴ One should note that it is a controversial issue whether the justification refers to the agent or to his act. On many (perhaps most) occasions these two ways of conceptualizing justification are equivalent but one can easily imagine a situation in which an agent’s reasonable but false belief (e.g., that someone wanted to kill him) justifies *him* (retrospectively, so to say), but not his act (since his act could not be treated as a permissible exception to some general, prohibitory rule).

⁵ Yet it should also be noted that the term “insanity defense” is misleading if it is to include (as it does) defense by reason of mental retardation.

to reconstruct the consequences of not only the philosophical-anthropological picture which I endorse, but also the consequences of alternative pictures.

The picture that underlies my analyses can be described in different ways: as the libertarian, the non-reductionist, the humanistic, or, most aptly, *the personalist*. It is based on the assumption that humans are truly unique beings, essentially different from animals. In consequence, any talk of non-human animals which implies that human beings are “human animals” is highly misleading as it overlooks this uniqueness and creates the false impression that the difference between humans and animals is one of degree. How can this uniqueness be described? Let me start from the observation that it is only humans that are not “closed” in their concrete perceptions (which can be called “the first-level representations of the world”). Thanks to their capacity to use abstract notions, they create abstract representations of the world (which can be called “the second-level representations of the world”). This capacity creates, as one may call it, an “ontological crevasse” between themselves and the world, and thereby makes them fundamentally “detached from” the world. Accordingly, a human being can direct his attention beyond his immediate (concrete) perceptions: either towards the “inside” – the depth of his self, or towards what lies “outside” his immediate perceptions. It is this capacity that Ernst Cassirer (1972) meant when saying that a human being is *animal symbolicum*. This capacity underlies (or is constitutive of) a human being’s most unique faculty – that of free will.

The faculty of free will can be defined in two equivalent ways, highlighting its different aspects. The first definition, which emphasizes its dispositional aspect, states that free will is autonomy – the capacity for self-determination in accordance with some hierarchy of values. The thought that human beings are capable of self-determination, that they are not determined by any “human nature,” but can choose their own nature, was memorably expressed by Pico della Mirandola in *Oratio de dignitate hominis*:

Statuit tandem optimus opifex, ut cui dare nihil proprium poterat commune esset quicquid privatum singulis fuerat. Igitur hominem accepit indiscretæ opus imaginis at que in mundi positum meditullio sic est alloquutus: “Nec certam sedem, nec propriam faciem, nec munus ullum peculiare tibi dedimus, o Adam, ut quam sedem, quam faciem, quæ munera tute optaveris, ea, pro voto pro tua sententia, habeas et possideas. Definita ceteris natura intra præscriptas a nobis leges coeretur. Tu, nullis angustiis coercitus, pro tuo arbitrio, in cuius manu te posui, tibi illam præfinies. Medium te mundi posui, ut circumspiceres inde commodius quicquid est in mundo. Nec te caelestem neque terrenum, neque mortalem neque immortalem fecimus, ut tui ipsius quasi arbitrarius honorariusque plastes et fîctor, in quam malueris tute formam effingas. Poteris in inferiora quæ sunt bruta degenerare; poteris in superiora quæ sunt divina ex tui animi sententia regenerari.” O summam Dei patris liberalitatem, summam et admirandum hominis felicitatem! Cui datum id habere

quod optat, id esse quod vellit ... Nascenti homini omnifaria semina et omnigenae vitae germina indidit Pater ... (Pico della Mirandola 2010: 35–44).⁶

Human nature is therefore “Prothean,” as Pico della Mirandola himself wrote: It can assume different forms, depending on a man’s decision. We can be what we want to be and so we are responsible for what we become. The second definition points at free will’s active (agential) aspect: An agent can be said to have free will if, and only if, when choosing a given action *a* at a given moment *m*, he is not causally determined to choose *this* action (*a*), i.e., can choose (without positing any change in the “state of the world” preceding the moment of choice *m*) also a different action (*not-a*) at the moment *m*. This aspect is captured by the famous definition of free will/free choice (*liberum arbitrium*) provided by Thomas Aquinas: *Liberum arbitrium est causa sui motus: quia homo per liberum arbitrium seipsum movet ad agendum* (*Summa Theologiae*, I, q. 83, a.1).⁷

It should be stressed that this – personalist – view of human nature cannot be conclusively proved. I shall confine myself to invoking two arguments in

⁶ English translation (by A.R. Caponigri):

At last, the Supreme Maker decreed that this creature, to whom He could give nothing wholly his own, should have a share in the particular endowment of every other creature. Taking man, therefore, this creature of indeterminate image, He set him in the middle of the world and thus spoke to him: “We have given you, Oh Adam; no visage proper to yourself, nor any endowment properly your own, in order that whatever place, whatever form, whatever gifts you may, with premeditation, select, these same you may have and possess through your own judgment and decision. The nature of all other creatures is defined and restricted within laws which We have laid down; you, by contrast, impeded by no such restrictions, may, by your own free will, to whose custody We have assigned you, trace for yourself the lineaments of your own nature. I have placed you at the very center of the world, so that from that vantage point you may with greater ease glance round about you on all that the world contains. We have made you a creature neither of heaven nor of earth, neither mortal nor immortal, in order that you may, as the free and proud shaper of your own being, fashion yourself in the form you may prefer. It will be in your power to descend to the lower, brutish forms of life; you will be able, through your own decision, to rise again to the superior orders whose life is divine.” Oh unsurpassed generosity of God the Father, Oh wondrous and unsurpassable felicity of man, to whom it is granted to have what he chooses, to be what he wills to be! ... Upon man, at the moment of his creation, God bestowed seeds pregnant with all possibilities, the germs of every form of life (6-9).

⁷ English translation (by the Fathers of the English Dominican Province): “Free-will is the cause of its own movement, because by his free-will man moves himself to act.”

its favor (as already mentioned, its deeper justification would be a topic for a different book).

The first argument states that if free will – the central faculty of the person – did not exist, there would be no moral justification for the practice of blaming and praising us for our actions. It must be admitted, however, that this argument can be criticized on the grounds that it presupposes the claim that the practice requires such a justification and that only the assumption of free will can provide it (I shall return to this argument in Chapter 1). The second one hints at the close connection between rationality and freedom of will: If we assume that human beings are capable of making rational judgments (this assumption is less controversial than the assumption of freedom of will), then we must agree that they are endowed with free will; the possessing of one of the capacities requires the possession of the other. These two arguments can be found in this classical passage from Thomas Aquinas's *Summa Theologiae*:

Homo est liberi arbitrii, alioquin frustra essent consilia, exhortationes, praecepta, prohibitiones, praemia et poenae. Ad cuius evidentiam, considerandum est quod quaedam agunt absque iudicio, sicut lapis movetur deorsum; et similiter omnia cognitione carentia. Quaedam autem agunt iudicio, sed non libero; sicut animalia bruta. Iudicat enim ovis videns lupum, eum esse fugiendum, naturali iudicio, et non libero, quia non ex collatione, sed ex naturali instinctu hoc iudicat. Et simile est de quolibet iudicio brutorum animalium. Sed homo agit iudicio, quia per vim cognoscitivam iudicat aliquid esse fugiendum vel prosequendum. *Sed quia iudicium istud non est ex naturali instinctu in particulari operabili, sed ex collatione quadam rationis; ideo agit libero iudicio, potens in diversa ferri.* Ratio enim circa contingentia habet viam ad opposita; ut patet in dialecticis syllogismis, et rhetoricis persuasionibus. Particularia autem operabilia sunt quaedam contingentia, et ideo circa ea iudicium rationis ad diversa se habet, et non est determinatum ad unum. *Et pro tanto necesse est quod homo sit liberi arbitrii, ex hoc ipso quod rationalis est* (*Summa Theologiae* I, q. 83, art. 1; emphasis added).⁸

⁸ English translation (by the Fathers of the English Dominican Province):
 Man has free-will: otherwise counsels, exhortations, commands, prohibitions, rewards, and punishments would be in vain. In order to make this evident, we must observe that some things act without judgment; as a stone moves downwards; and in like manner all things which lack knowledge. And some act from judgment, but not a free judgment; as brute animals. For the sheep, seeing the wolf, judges it a thing to be shunned, from a natural and not a free judgment, because it judges, not from reason, but from natural instinct. And the same thing is to be said of any judgment of brute animals. But man acts from judgment, because by his apprehensive power he judges that something should be avoided or sought. *But because this judgment, in the case of some particular act, is not from a natural instinct, but from some act of comparison in the reason, therefore he acts from free judgment and retains the power of being inclined to various things.* For reason in contingent matters may follow opposite courses, as we see in dialectic syllogisms and rhetorical arguments. Now particular operations are

In my view, the second argument is very cogent: Human rationality indeed cannot be separated from the power of will and its freedom. One could object to this claim by saying that one can easily imagine a rational agent, that is: responsive to reasons, capable of weighing and comparing them, of deliberating about ends to pursue, and of selecting proper means to realize these ends, who is not free. Indeed, it *might seem* that a rational agent could act deterministically: The whole process of perceiving reasons, grasping them, and acting on them could proceed according to inexorable causal laws. Yet, upon deeper reflection, this objection proves to be implausible – vulnerable to two serious counter-objections. The first counter-objection, in its weaker form, states that it cannot be claimed at the same time that man is *animal rationale* (whose fundamental distinction is thinking, which is an intentional and teleological process) and that materialistic naturalism (according to which thinking is essentially a causal-mechanical process) is true. Consequently, we have to choose between the image of man as *animal rationale* and materialistic naturalism. In its stronger form, the counter-objection not only points to the necessity to choose between materialistic naturalism and the image of man as *animal rationale*, but it also shows that materialistic naturalism is an inconsistent, self-defeating view. More specifically, it shows that although the naturalist's statement "materialistic naturalism is true" is not formally incoherent (i.e., is not internally contradictory), it is self-referentially incoherent, i.e., undermines itself, because in the light of what it preaches (materialistic naturalism), it is just the result of a series of causally connected events occurring in the brain. This gives rise to the obvious question: Why should we treat this statement seriously if – ultimately – its origins have nothing to do with reasons, i.e., with arguments put forward by the free mind to support a certain conviction?⁹ The second counter-objection states that an essential feature of a rational agent is consciousness and self-consciousness. In fact, this feature is built into the above definition of rationality: It is hard to imagine that the very process of weighing (attaching weights to) reasons, comparing them, especially (which happens frequently) when these reasons are inconclusive (i.e., when they have this kind of character which excludes applying to them formal methods – rules of mathematics, logic, or probability), and grasping and being guided by them could proceed without the active participation of consciousness. And consciousness can hardly be separated from the will: The final decisions of what weight to assign to various reasons for action and what action to take belong

contingent, and therefore in such matters the judgment of reason may follow opposite courses, and is not determinate to one. *And forasmuch as man is rational is it necessary that man have a free will* (emphasis added).

⁹ I develop this line of argumentation at great length in Załuski 2017.

to the prerogative of the will. This last point can be developed along the lines indicated by John Searle (2010: 123), who wrote about the subjective experience of three “gaps”: between reasons for an action and a decision to take this action, between a decision what to do and performing the action, and between starting an action and carrying it through to a conclusion. In Searle’s view, these three gaps “are all parts of a continuous causal gap of conscious voluntary action” (2010: 123) – they are the experiences “of thoughts and actions where we do not sense the decisions and actions as causally fixed by the antecedent conditions” (2010: 123). Obviously, one could try to argue that the gaps are not real but illusionary – and their illusion results from our insufficient knowledge of the antecedent causes which deterministically forejudge what decision and action is taken. But even though this objection cannot be falsified (one can always posit the existence of some unknown causes to be revealed by the future development of science), it is highly implausible. It remains in deep tension with our common intuitions (which strongly suggest – against determinism implying that the physical world is closed to any non-physical influences – that our consciousness, understood in a non-reductionist manner, is causally efficacious). It would also lead to the undermining of the very idea of human agency. Thus, determinism arguably has insurmountable difficulties in accounting for the very existence of consciousness and of the related fact of human rationality (humans’ ability to grasp and be guided by reasons).

As already mentioned, the above-sketched picture of human nature can be called “humanistic.” In describing it this way, I invoke one of the several meanings of the word “humanism,” viz. as referring to the view which asserts that human beings occupy a distinguished place in the universe among other living beings owing to the fact that they are autonomous (in the sense that they have free will) and rational (in the sense that they are responsive to reasons and act on reasons, or at least have the capacity to act on reasons). Therefore humanism, at least in the Renaissance meaning of this term, rejects the claim that human beings are nothing more than physical systems whose behavior is strictly governed by deterministic laws (some additional aspects of humanism will be highlighted below). Since the agent who satisfies both conditions – freedom of will and rationality – is called “a person,” this view of human being can also be called “personalist.” I shall use this last term throughout the rest of this book since it lacks the ambiguity that the term “humanism” is burdened with; the latter is at times used also by scholars who reject one (or both) of the assumptions that are crucial for the classical account of the person (the account developed, for instance, by Thomas Aquinas, Emmanuel Mounier, or Karol Wojtyła/John Paul II, and adopted in this book), viz. that human beings have free will; and that there exist objective values which human beings strive

to realize.¹⁰ By way of summary, let me invoke the famous definition by Emmanuel Mounier, according to which a person is

un être spirituel constitué comme tel par une manière de subsistance et d'indépendance dans son être; elle entretient cette substance par son adhésion à une hiérarchie de valeurs librement adoptées, assimilées et vécues par un engagement responsable et une constant conversion; elle unifie ainsi toute son activité dans la liberté et développe par surcroît, à coups d'actes créateurs, la singularité de sa vocation (Mounier 1936: 63).¹¹

This is not a complete picture of the personalist view but sufficiently detailed at this stage of my considerations. But much more needs to be said about the will – as part of this picture. This will be done in Chapter 3, prior to making an analysis of the volitional component of the insanity defense. In summary, *it is an essential feature of the personalist view that it defines human being first and foremost in terms of reason and (free) will – his spiritual powers, rather than in terms of desires or instincts – his affective and instinctual side.*

Let me repeat that my analyses of the insanity defense will be more than just a matter of deriving conclusions from the above philosophical-anthropological assumptions. Above all, they will have a contextual character: I shall present my view of the insanity defense in the broader context of other views, based on different (philosophical-anthropological) assumptions (which I shall be presenting, though not in a systematic manner, throughout this

¹⁰ The first assumption is rejected by what can be called *naturalistic* “humanism,” having little in common with the original version of humanism defended, for example, by Pico della Mirandola. A good example is Michael Schmidt-Salomon’s (2012) book *Manifest des evolutionären Humanismus*, in which he develops a doctrine he calls “evolutionary humanism.” He understands this kind of humanism only in purely normative terms as a postulate of the “humanization” (improvement) of the living conditions of human beings. He does not regard the assumption of freedom and rationality as components of humanism. The view of man assumed by Schmidt-Salomon is in fact purely naturalistic. One can discern a related trend in understanding humanism among those thinkers who are motivated by the fascination of the technological achievements of human beings (their version of humanism can be called “Promethean humanism”). The second assumption was rejected by “existentialist humanism”; its representatives (e.g., Jean Paul Sartre) conceded that human beings are endowed with free will but denied the existence of objective values; in their view, values are not discovered but created by human beings.

¹¹ Translation (by the author):

a spiritual being constituted as such by his way of subsistence and independence of his being; he upkeeps this substance by his adherence to a freely adopted hierarchy of values, assimilated and lived by a responsible engagement and constant conversion; he unifies in this way his whole activity in liberty and, additionally, develops, by creative acts, this singularity of his vocation.

book). As already mentioned, this approach to the question of the insanity defense is based on the premise that there is a connection (a “bridge”) between philosophical-anthropological assumptions and the views on the insanity defense. But how might one describe this connection in a more precise manner? The answer to this question is straightforward: The connection goes *via* conceptions of (moral and legal) responsibility. In other words, a given picture of human beings implies a certain conception of responsibility which, by itself or with some additional premises (especially, as we shall see, the premises determining the view of mental illness), influences the way in which we understand the responsibility of the mentally ill (insane). It is therefore pertinent to clearly indicate here what conception of responsibility is implied by – or is most naturally connected to – the personalist picture of human beings assumed in this book.

Most generally, it can be said that the personalist picture implies that free will is a necessary condition of responsibility (of a person *P* for committing an offence *a*). However, this description is somewhat vague as it allows three quite different “free will-based” conceptions of responsibility (i.e., such conceptions which imply that human beings have free will, and that free will is at least a necessary condition of responsibility).

According to the first one (the stricter), a person *P* is responsible for his action *a* if, and only if, *P* has free will and his action was undertaken in the absence of external coercion (of the so-called *vis compulsiva* or *vis absoluta*).¹² The second (less strict) states that a person *P* is responsible for his action *a* if, and only if, *P* has free will and, in undertaking this action, he was free from both external and internal coercion. These two conceptions imply that responsibility is not gradable: Its ascription is an all-or-nothing question. The first conception is stricter because it assumes that free will is “activated” even among people whose actions are not fully voluntary, i.e., are not free from internal coercion (e.g., among the mentally ill). Let me note, *en passant*, that it is precisely this conception which was assumed by many of the adherents of the abolition of the insanity defense in some “conservative” states in the US (e.g., Idaho, Montana). The picture of human nature which they endorsed can be called *radically libertarian*: It assumes that free will is never “suspended” among the agents, and, consequently, it is as effective among the mentally ill

¹² An “action” done as a result of *vis absoluta* (one of two forms of external coercion), in which the decisional component is entirely absent (because an agent is physically forced to take a given “action”), is not an act in the strict sense. In the case of *vis compulsiva* (another form of external coercion) where the agent is faced with a threat and thereby coerced into choosing a “lesser evil,” the decisional component is present. But he is forced to act, so to say, against himself, so responsibility ascription would be quite controversial in such a case (although the dispute on this issue is still ongoing).

as among the mentally sane. This conception is *radically libertarian* but it can hardly be called personalist: A truly personalist conception assumes a *less rigid* account of free will; *it stresses the uniqueness of human beings, but does not deny the fact that their unique faculties may sometimes fail to function properly* (this is another aspect of personalist humanism, which supplements those mentioned above). Thus, the second conception of responsibility is *personalist*: It prohibits blaming an agent for his actions which are not fully voluntary (i.e., which are not free from *both* external and internal coercion). This conception can be understood in two slightly different ways: We can say either that it implies that free will is “activated” only among those whose actions are fully voluntary (free from internal and external coercion), or that free will is always “active” (even among those whose actions are not fully voluntary) but it can ground an ascription of responsibility *only if* an action is fully voluntary.

Apart from the two above-mentioned “free will-based” conceptions, one can also distinguish the third one, which implies that (moral and legal) responsibility is a matter of degree: *It is proportional to our moral desert*, to what we do, thanks to our free will, with our character – our capacities and propensities that were “given” to us (by our genes or the environment), also taking into account the historical circumstances in which we act. This conception requires that we take some forms of moral luck into account when making ascriptions of responsibility, viz. character luck and situation luck. This conception, which is also (like the second one) justified by the personalist view of human beings,¹³ was elegantly described by Clive Staples Lewis in the following passage:

The bad psychological material is not a sin but a disease. It does not need to be repented of, but to be cured. And by the way, that is very important. Human beings judge one another by their external actions. God judges them by their moral choices. When a neurotic who has a pathological horror of cats forces himself to pick up a cat for some good reason, it is quite possible that in God’s eyes he has shown more courage than a healthy man may have shown in winning the V.C. When a man who has been perverted from his youth and taught that cruelty is the right thing, does some tiny little kindness, or refrains from some cruelty he might have committed, and thereby, perhaps, risks being sneered at by his companions, he may, in God’s eyes, be doing more than you and I would do if we gave up life itself for a friend. It is as well to put this the other way round. Some of us who seem quite nice people may, in fact, have made so little use of a good heredity and a good upbringing that we are really worse than those whom we regard as fiends. Can we be quite certain how we should have behaved if we had been saddled with the psychological outfit, and then with the bad upbringing, and then with the power, say, of Himmler? That is why Christians are told not to judge. We see only the results which a man’s choices

¹³ Though, it should be added that this conception is somewhat in tension with the radical understanding of self-determination, typical, for instance, of the above-quoted Pico della Mirandola.

make out of his raw material. But God does not judge him on the raw material at all, but on what he has done with it. Most of the man's psychological makeup is probably due to his body: when his body dies all that will fall off him, and the real central man, the thing that chose, that made the best or the worst out of this material, will stand naked. All sorts of nice things which we thought our own, but which were really due to a good digestion, will fall off some of us: all sorts of nasty things which were due to complexes or bad health will fall off others. We shall then, for the first time, see every one as he really was. There will be surprises (Lewis 2009: 91–92).

I apologize to the reader for this long quote, yet it perfectly encapsulates the spirit of the conception that will play some role in my further analyses (especially when dealing with the problem of responsibility of psychopaths in Chapter 2). This conception, which I term the “cosmic/God's eyes conception of responsibility,” can be more precisely presented in the following way: A person *P* is responsible (blamable) for his action *a* if, and only if, *P* has free will and his action was fully voluntary (free from internal and external coercion), and his responsibility (blame) is a matter of degree: The easier it was for him (given his “psychological makeup”) to refrain from doing *a*, the higher the degree of his responsibility for *a*.¹⁴ Thus, according to this conception, what is of crucial importance is not the result or achievement but the moral effort to reach the result or achievement – the fact an agent has done the best he can with what he has at his disposal (his natural abilities and propensities).

Let me summarize: All three conceptions of responsibility are libertarian (in the sense that they are based on the assumption that free will is not an illusion), but only the last two can plausibly be called personalist. The question is: Which of these three conceptions is most plausible? The first one is *too* strict: It leads to blaming people who arguably should not be blamed (those whose actions were not fully voluntary); furthermore, it relies on the dubious descriptive assumption that free will functions in the same way among those whose actions are fully voluntary and those whose actions are not. Consequently, I assume that the second conception is an acceptable conception of moral and legal responsibility. It provides both a deep justification of punishment and, at the same time, can be practically realized by human beings. This last feature distinguishes it from the third conception, which may be, all things considered, the *right one* (as it arguably treats the idea of moral desert in a nuanced and proper way) but it cannot be realized by humans (hence the name I have given it here: “the cosmic/God's eyes”).

¹⁴ It seems that this conception, putting so much stress on moral desert/effort, could not have arisen without the influence of Christianity: In ancient ethical thought, the role of desert was not emphasized or marginalized (which was probably due to the fact that ancient thinkers did not know, or more precisely did not discover, the faculty of free will, without which the concept of moral desert loses its philosophical foundations).

What I have described so far are the two presuppositions of my analyses of the insanity defense, viz. the personalist picture of human beings and (the consequent) “free will-based” conception of responsibility, which states that free will *plus* full voluntariness are the necessary and jointly sufficient conditions of moral and legal responsibility. But how, in detail, will these analyses be carried out in this work? Their structure will be dictated by the fact that the insanity defense can be examined at two different levels. The first level concerns its most general assumption, viz. that the mentally sane can be held criminally responsible, and the mentally ill cannot (I shall call it the “Standard Model” of the criminal responsibility of the mentally ill). This assumption is common for all specific varieties of the insanity defense, which are distinguished according to a different criterion, viz. according to the *psychological content* of mental incompetence, which they recognize as releasing the accused from responsibility. The content may be narrow (thin) or broad (thick): It may embrace the cognitive component alone (“I did not know what I was doing,” “I did not know that what I was doing was wrong”) or also the volitional one (“I could not control my action, I could not refrain from making this action”), and each of these components can, in turn, be understood narrowly – less leniently/more restrictively – or broadly – more leniently/less restrictively. The narrow understanding of the structure of the insanity defense (confining it only to the cognitive component) and the narrow understanding of any of its components (excluding, e.g., “affective knowledge” from the cognitive test) are more restrictive (less lenient) towards offenders because they amount to providing less demanding conditions for responsibility ascriptions: They have the effect of heightening the threshold for a successful insanity defense and, consequently, of decreasing the number of offenders who can effectively invoke it during a trial (the threshold is highest if the structure of the defense is reduced to its cognitive component and this component is understood narrowly, and lowest if the defense includes both components understood broadly). Now, it is precisely the psychological content that constitutes the second level of analysis of the insanity defense. It is worth noticing that the first level is rarely distinguished in the relevant literature: The assumption that underlies the insanity defense – that only the mentally sane can be held morally responsible – usually appears to be taken for granted, and thereby not analyzed at great length if it is analyzed at all. But such an approach impoverishes the analysis of the insanity defense. It overlooks the fact that the criminal responsibility of the mentally ill can be – and in some jurisdictions *is* – regulated in a different way: Mental insanity is not always a legal excuse releasing the offender from criminal responsibility. Furthermore, even if all criminal justice systems assumed the insanity defense, it would still be necessary, if we wished to gain better understanding of it, to carefully trace its philosophical-anthropological assumptions

(as well as the assumptions of the hypothetical other solutions to the problem of the criminal responsibility of the mentally ill).

These two levels of the analysis of the insanity defense can be easily discerned in the structure of this book. The task of identifying the philosophical assumptions of the Standard Model and other models of criminal responsibility of the mentally ill is undertaken in Chapter 1. I distinguish the models on the basis of two criteria: the mental state of the offender (mentally sane or mentally insane), and criminal responsibility (admitted or excluded). As a result, four main models are distinguished (variants of them are also analyzed in Chapter 1):

(M1 – *Standard*): Mentally sane → Responsible, Mentally insane → Not-responsible

(M2 – *Perverse*): Mentally sane → Not-Responsible, Mentally insane → Responsible

(M3 – *Negative Equalization*): Mentally sane → Not-Responsible, Mentally insane → Not-Responsible

(M4 – *Positive Equalization*): Mentally sane → Responsible, Mentally insane → Responsible

I shall try to provide a thorough analysis of their philosophical underpinnings, arguing that there are two main factors which determine the choice among them – the assumed conception of responsibility and the account they provide of mental illness. Accordingly, I shall supply a detailed analysis of various conceptions of responsibility (in addition to those which have been presented above) and of two accounts of mental illness (which I call “affirmative” and “negationist”). The analysis of these accounts is not purely theoretical; it also draws considerably on the history of 20th-century psychiatry, with a particularly detailed presentation of the anti-psychiatry of the 1960s. It was within this current that the “negationist” account was developed, and this work will examine its influence on the shaping of the institution of the insanity defense in various legal systems. In this chapter I advocate the already mentioned “free will-based” conception of criminal responsibility and an affirmative view of mental illness, which jointly justify the Standard Model of the criminal responsibility of the mentally ill, i.e., the insanity defense. Chapter 1 therefore answers the question of why the law should recognize the insanity defense at all. However, it does not say anything about the concrete shape that this type of defense should take, i.e., whether it should contain both components – the cognitive and the volitional – and, if so, how they should be understood (narrowly or broadly). This second level of the analysis of the insanity defense is developed in the remaining two chapters.

In Chapter 2, I first present various possible formulations of the cognitive component, drawing on the history of the insanity defense in various legal systems. I then pass to the analysis of the very notion of normative (moral and legal) knowledge (crucial for this component). This will lead me to the problem of the moral and legal responsibility of psychopaths. I formulate a critical analysis of various solutions to this problem proposed in the relevant literature and reject the following: that psychopaths can be held legally responsible but not morally responsible; that they cannot be held responsible either morally or legally; that they cannot be held morally responsible and can be held legally responsible but only for the so-called *mala prohibita*. I defend the following solution: Psychopaths can be held responsible both morally and legally (like all “normal,” i.e., non-psychopath, offenders), and their specific features (e.g., their lack of – or at least deficits in – moral emotions and empathy) can only matter when we shift our ethical perspective and invoke the “cosmic/God’s eyes” of moral responsibility. As mentioned before, I find two “free will-based” conceptions (the non-gradable one and the gradable one – the “cosmic/God’s eyes”) of moral responsibility plausible but in different contexts: While the first one is suitable for human moral judgment, the other one involves some “external” (supra-human, God’s), more objective moral judgment as it is based on information which, in principle, is inaccessible to human beings. If we wanted to dispense with metaphysics or theology and limit ourselves to human moral judgment alone, we could say that the first conception must be invoked when we make “normal” moral judgments, while the second one may be proper in those moments in which we try to engage in a deeper (“cosmic/God’s eyes”) moral reflection. The typical result of this kind of deeper moral reflection is likely to be a stronger propensity to forgiveness since this reflection must be accompanied by ethical and intellectual humility, flowing from the awareness of our own sinfulness and from our limited access to information about the agents; this humility should lead to our interpreting all uncertain information in favor of the person being morally judged; only God or a “cosmic observer,” knowing all the relevant information, will have the right to be a severe judge. But my claim that the specific features of psychopaths should be taken into account while “switching” to the “cosmic/God’s eyes” conception of responsibility does not imply that their case is unique: The degree of the responsibility of all agents, whether psychopathic or not, will have to be properly accommodated to their “psychological makeup.” It may, indeed, turn out that the features of psychopaths substantially decrease their responsibility; but we (as human beings) cannot know that. Equally it may emerge that the “psychological makeup” of many non-psychopathic agents is a more serious obstacle to moral behavior than that of psychopaths (although, let me repeat: If *human beings* engage in a moral reflection based on the “cosmic/God’s eyes” conception of responsibility the typical results

should be an increased readiness to forgiveness). In the course of my analysis of the responsibility of psychopaths I shall also examine the problem of the borderline between madness and extreme badness (evil), trying to answer the question of whether extreme evil can be imputed to agents or whether it should be regarded as a manifestation of their madness. I will distinguish five different senses of extreme evil: (1) doing evil just for the sake of doing evil; (2) sadistic wrongdoing, i.e., deriving pleasure from making the other person suffer; (3) doing evil for no intelligible reason; (4) doing non-natural evil, that is, violating the most elementary norms presumably inscribed in our nature by biological evolution, e.g., norms prohibiting us from killing our parents, siblings, children; and (5) doing beastly or brutish evil (in the sense of the Aristotelian *theriotes*) – evil that is absolutely horrifying for it resembles the actions of wild beasts. I shall argue that only the last three types of extreme evil, and especially type (3) and (5), can be regarded as madness, which means that there are types of extreme evil which can be committed by mentally sane people. Therefore, extreme evil, *ipso facto*, does not amount to madness.

In Chapter 3 I provide a critique of the volitional component, a critique which will have several layers. The most general one is connected with the personalist picture of human nature adopted in this book. The picture, as already mentioned, assumes that human beings are endowed with autonomy – the capacity for self-determination. This implies that they possess a certain faculty – will – and this faculty is understood as capable of resisting various internal and external influences. This understanding of will, assumed within the so-called classical tradition of philosophy and constituting an essential element of the personalist picture of man, implies that will is a *decisive* faculty of action; if it is “defeated,” i.e., if an agent acts contrary to his will, then he can be blamed for it. The chapter provides a more detailed account of this understanding of will from a historical perspective, along with support for the claim that this personalist philosophical-anthropological conception casts more than a shadow of doubt on the volitional component of the insanity defense. However, my critique of this component is not only based on this conception. I formulate several other, more specific, objections against the inclusion of the volitional component into the insanity defense; I term them as follows: “the epistemic objection,” “the non-existence objection,” “the slippery slope objection,” and finally, “the formal justice objection.” This critique will not lead, however, to the rejection of the volitional component but rather to the postulate of its very restrictive or less lenient understanding. The chapter also traces the further implications of the critique of the volitional component of the insanity defense, viz. those regarding the so-called crimes of passion. I argue, drawing on my critical analysis of the volitional component of the insanity defense, that there are no cogent reasons to distinguish them as a separate category of crimes that deserve more lenient treatment.